LEVEL II

Research Memorandum 78-7

CONTENT ANALYSIS AND THE ORGANIZATION OF COMBAT INTELLIGENCE DATA

Murray S. Miron and Samuel M. Patten
Syracuse University Research Corporation

and

Stanley M. Halpin
Army Research Institute for the Behavioral and Social Sciences

BATTLEFIELD INFORMATION SYSTEMS TECHNICAL AREA

U. S. Army
Research Institute for the Behavioral and Social Sciences

February 1978

Army Project Number
16 2Q762101A754

Intelligence Systems

9 Research Memorandum 78-7

6 CONTENT ANALYSIS AND THE ORGANIZATION OF COMBAT INTELLIGENCE DATA.

10 Murray S. Miron and Samuel M. Patten
Stanley M. Halpin

12 48

Stanley M. Halpin
Army Research Institute for the Behavioral and Social Sciences

Submitted by:
Edgar M. Johnson, Chief
Battlefield Information Systems Technical Area

11 February 1978

14 ARI-RM-78-7

Accession For	
NTIS GRA&I	X
DDC TAB	
Unannounced	
Justification	
By	
Distribution/	
Availability Codes	
Dist	Avail and/or special
A	

Approved by:

Joseph Zeidner, Director
Organizations and Systems Research Laboratory

J. E. Uhlaner, Technical Director
U.S. Army Research Institute for the Behavioral and Social Sciences

---

Research Memorandums are informal reports on technical research problems. Limited distribution is made, primarily to personnel engaged in research for the Army Research Institute.

CONTENT ANALYSIS AND THE ORGANIZATION OF COMBAT INTELLIGENCE DATA

INTRODUCTION

In order to expeditiously and accurately process intelligence information, it is essential that incoming information be systematically organized. Such organization provides a means for categorizing, differentiating and integrating intelligence for retrieval, evaluation, and interpretation. This report will examine and test an application of a high-speed data processing technique which is designed to automatically provide organizational structure for incoming intelligence. The procedure involves the use of a system of computer routines known as the General Inquirer, which was developed for the analysis of message content. The routines, originally devised by a team of researchers at Harvard University (see Stone et al., 1962) have been modified and upgraded extensively by J. Philip Miller of St. Louis University (Miller and Psathas, 1968) and to a lesser extent by the senior author.

The approach to the organization of intelligence information represented by these computer procedures involves the automatic identification and cataloging of a set of previously selected word and phrase forms in the text of the intelligence reports received from the field. Critical word occurrences in the messages are organized into a set of concept categories flexibly defined by the intelligence analyst. This flexibility, which is essential if the system is to be responsive to the specifics of either particular situations or the particular needs of any given analyst, is achieved through the use of a series of user-oriented programs which employ a syntax readily mastered by the analyst for specifying identification and retrieval operations. Thus, although the programs are internally quite complex, from the standpoint of the analyst-user they are simple to use, allowing for a wide range of options which place all of the computational burden on the computer.

At the time that the General Inquirer was developed, most computer techniques were such that numeric processing was comparatively easy and text processing difficult. The General Inquirer was a much-needed tool in fields that dealt with textual data. The original authors described it as:

> ...a set of computer programs to (a) identify systematically, within text, instances of words and phrases that belong to categories specified by the investigator; (b) count occurrences and specify co-occurrences of these categories; (c) print and graph tabulations; (d) perform statistical tests; and (e) sort and regroup sentences according to whether they contain instances of a particular category or combination of categories (Stone et al., 1966, p. 68).

In content analysis, the General Inquirer functions as a well-trained clerk who assigns particular categories (specified by the investigator before the analysis) to words and/or combinations of words. The IBM S/360-S/370 versions of the General Inquirer, called the Inquirer II (abbreviated I/II) still contain this capability and also allow for more elaborate analyses of the data. The Inquirer II programs are able to make more elaborate searches of the textual data and provide more options to the potential user.

Content analysis may be defined as a data organization technique which involves a systematic identification of theoretically relevant categories in textual data. As employed in this project, the procedure provides a method for deriving a taxonomy of intelligence reports. Categories of report content and the rules by which they may be identified are defined by means of a dictionary. For the purposes of this project, a special purpose dictionary was constructed from analysis of intelligence reports taken from the intelligence journals of the 28th Infantry Division for the 10-15 December 1944 period just before the German Ardennes counteroffensive known as the Battle of the Bulge. This dictionary provides a taxonomy of the content of the military messages; the procedure described below then organizes, integrates and classifies the messages on the basis of their content similarities and differences. Order is imposed on the otherwise unorganized reports through the identification of dictionary-defined concepts in the reports. Reports which polythetically share the greatest numbers of concept occurrences will be considered "similar," and separated from those not sharing such occurrences of concepts.

Such an approach to the organization of intelligence data allows the individual analyst to flexibly define his own categories of message content and structure. Additionally, it allows for continuous updating and modification of the organizational schema as the situation requires. In a field application of such a system, the sequentially received reports would be entered into the computer according to the conventions outlined below and successive factor structures would be computed for the body of reports forming the data base at any given time. As each successive report is added to the data base, or at any other appropriate time, e.g., at the end of the day, a new factor structure and organization would be computed.

This report presents a factor structure on the basis of an analysis of 40 reports over six days; this represents the covergent outcome of what would have been a succession of structures computed after six days of reports had been received. As these intermediate structures converge, many of the category variables would not be employed. Estimates of the efficiency of the categories and these intermediary structures may be made. Category definitions in the dictionary may be modified and updated by the analyst, as required, to produce an organizational structure which he deems to be meaningful and which at the same time efficiently, economically, and successfully tags the content of the reports. For such purposes, the analyst would inspect the untagged word file to ascertain which

items should be added to the dictionary and then would produce a new structure. Over a period of time, each analyst or installation would thus build a dictionary of intelligence concepts which would be uniquely suited to the type of material and situations being analyzed.

METHOD

The general procedure for the analysis was: (1) a sample of 40 intelligence reports was keypunched into IBM cards according to a set of conventions, (2) a dictionary of critical concepts was constructed, (3) rules for identification of these concepts as they occurred in the text were developed, (4) tabulations of the occurrences of the identified concepts were calculated for each message, and (5) correlations and factor analyses were computed using these tabulations of identified concepts. Full details of the syntax and computer routines employed are in the Appendix.

The intelligence reports (see Miron, Patten and Halpin, 1977) and the specially constructed dictionary served as inputs to the computer system. The program which assigned the categories (tagging program) read in the data a sentence at a time, then located each word in the dictionary. Instructions were given by the dictionary as to what category should be assigned and/or what searches of the context in which the word occurred should be made. The instructions were then executed. When the analysis of a sentence was completed (i.e., all the categories to be assigned had been assigned and all searches had been completed), the tagging program wrote out that sentence and read in the next. The process continued a sentence at a time until all the reports had been tagged. The output from the tagging program was a tagged file which was stored so that retrieval, tabulation, and statistical analyses of the data could be made.

DATA INPUT AND FORMAT

Thirty-three intelligence reports actually received by the 28th Infantry Division, and seven false reports designed to test the capabilities of the system, served as the data base in the present study. These intelligence messages range in content from trival sighting reports of horsedrawn vehicles to G2 summaries of considerable tactical and strategic importance. The procedures outlined here may be followed with any sample of messages, without restriction as to the type or source of the message. However, it is expected that more meaningful report taxonomies will be obtained if each message covers a relatively limited scope of information.

The report data input is prepared as continuous text as if it were being typed. Each different character of the input text is assigned a function. For example, an alphanumeric character (A-Z, a-z, 0-9) is considered part of a word and a blank indicates the end of a word. The period (.), exclamation point (!), and question mark (?) indicate the end

of a sentence. Braces ([ ]), greater than and less than signs (> <), and the dollar sign ($) indicate message identification, titles, and comments which are not to be searched for in the dictionary and not given a content category.

DICTIONARY PREPARATION

The major task in using the Inquirer II system is the creation of a dictionary. A content analysis category (called a concept in the Inquirer systems) consists of a number of language signs (such as words, idioms, and phrases) which together represent a variable in the investigator's theory. For example, the analyst concerned about the movements of a particular enemy division might be interested in identifying the number of references to that division in reports already received and therefore constructs a category; e.g., "the 26th Volksgrenadier Division," composed of references to that division (Volksgrenadier, VG, 26th VG Division, 77th VG Regiment, 78th VG Regiment, etc.). The basic procedure in content analysis is to identify (tag) these signs when and if they occur in the text as instances of a particular concept, and score them as such.

The analyst would seldom carry out a content analysis with a single concept. Instead, he is usually interested in examining relations of a number of semantic categories as they appear in intelligence documents. Therefore, we use a cluster of concepts, referred to as a content analysis dictionary. For the Inquirer system the exposition of this dictionary is in a special language, Dictionary Definition Language (DDL). Details of the syntax of this language are in the Appendix.

Category Construction. The first task in dictionary construction is to define the categories or concepts which were to be identified in the reports under consideration. The listing of these categories, as well as preliminary conceptual definitions of each of the concepts, forms an important nucleus for the actual construction of the dictionary. From this definition of the concepts and their interrelationships, the Concept Name Paragraph (CNP) is constructed. The Inquirer System allows an analyst the flexibility of assigning several different types of interrelations between concepts. The first of these indicates a one-to-one relation between the concept and the tags. In the Inquirer System concepts have names and tags have numbers. The concepts are used in constructing the entry used in the dictionary and at post-processing time for tabulations and listings; during the tagging and searching, it is the tag numbers which are used.

Table 1 displays the Concept Name Paragraph (CNP) or outline structure of the military dictionary devised for this project. These concepts, in their subordinate levels of organization, are indicated by successive subdividing strings of number identifiers. Thus, in Table 1, AREA OF OPERATIONS (1) is subdivided into locations (1,2), terrain (1,17), and

urban (1,18) with subcategories for each, e.g., as in this case, coordinates (1,2,3) as a subdivision of locations (1,2) and even further subdivision of the coordinates into sectors along the forward edge of the battle area. The methodology allows the analyst to add to or subtract from the dictionary, to reorganize categories, or to change the entries.

The concept categories used in the present study evolved from our earlier attempts to produce a subjective taxonomy of intelligence information (see Patten, 1974 and from a detailed examination of the Key Word in Context (KWIC) output listings of the reports themselves (Figure 1). No brief is made that this conceptual structure is either definitive or exhaustive; the dictionary is presented simply as a part of this demonstration of the methodological approach. However, as will be seen, the empirical test of this dictionary does produce a practical taxonomy of the reports on which it was tested.

Entry Selections. After the preliminary definition of the concepts and their interrelations have been completed, the next task is to determine what entries are to be in the dictionary. Two separate philosophies appear at this point. One is that nearly every word in the intelligence reports to be analyzed should be in the dictionary. This philosophy of an exhaustive dictionary has certain methodological attractions, chief among which are that the analyst has considered every possible word and account of those words which have not been found or tagged in the dictionary. This provides some measure of adequacy of the dictionary. Previous research has tended to show that the dictionary in the 3,000 to 4,000 word range will tag somewhere between 90 and 98% of ordinary texts. The other philosophy is that of a selective dictionary in which only words which are relevant to the concepts at hand are included. In most cases with the selective dictionary, it is possible to determine exhaustive lists of words which are to be assigned a given concept. This is the approach which has been taken in this project. For example, the concept COORDINATES exhaustively categorizes the locus of any action in the AREA OF OPERATIONS. Similarly, ORGANIZATION includes all organizational sub-divisions encountered in the message sample.

The dictionary underwent a number of revisions before reaching its final form. These revisions were made after inspecting the listings of untagged words and the KWIC outputs. In addition, a logical taxonomy of intelligence information was constructed according to a set of subjective procedures (Patten, 1974). In brief, the dictionary construction procedures seek to group significant word occurrences in the report sample under a set of major and minor category divisions of roughly equal scope. The total process is selective in that it uses only those items which suggest logically derived categories.

Entry Name Paragraph (ENP). Once the list of potential entry words is constructed, the relation between entry words and the concepts is specified. For this purpose, an exhaustive listing of the vocabulary of the reports may be made by means of the Key Word in Context (KWIC) routine (Figure 1). Where more than one word sense appears in a portion of text

Table 1

Concept Name Paragraph (CNP) of the Military Dictionary

AREA OF OPERATIONS = 1.
locations = 1, 2.
coordinates = 1, 2, 3.
north west = 1, 2, 3, 4.
north east = 1, 2, 3, 5.
north central west = 1, 2, 3, 6.
north central east = 1, 2, 3, 7.
south central west = 1, 2, 3, 8.
south central east = 1, 2, 3, 9.
south west = 1, 2, 3, 10.
south east = 1, 2, 3, 11.
zone = 1, 2, 12.
zone north = 1, 2, 12, 13.
zone north central = 1, 2, 12, 14.
zone south central = 1, 2, 12, 15.
zone south = 1, 2, 12, 16.
terrain = 1, 17.
urban = 1, 18.
BRANCH = 19.
armor = 19, 20.
artillery = 19, 21.
infantry = 19, 22.
CHANGE = 23.
decrease = 23, 24.
increase = 23, 25.
DIRECTION = 26.
north = 26, 27.
east = 26, 28.
south = 26, 29.
west = 26, 30
EXTENT = 31.
large = 31, 32.
small = 31, 33.
INTELLIGENCE = 34.
cognition = 34, 35.
reconnaissance = 34, 36.
LOGISTICS = 37.
MOVEMENT = 38.
ORGANIZATION = 39.
army = 39, 40.
army air corps = 39, 41.
corps = 39, 42.
division = 39, 43.
regiment = 39, 44.
battalion = 39, 45.
company = 39, 46.
headquarters = 39, 47.
team = 39, 48.

Table 1 (Cont'd)

Concept Name Paragraph (CNP) of the Military Dictionary

ORGANIZATION STATUS = 49.
- friendly = 49, 50.
- enemy = 49, 51

PERSONNEL = 52.
- civilian = 52, 53.
- military = 52, 54.
  - deserters = 52, 54, 55.
  - enlisted = 52, 54, 56.
  - officers = 52, 54, 57.
  - POWs = 52, 54, 58.

PLANES = 59.
- bombers = 59, 60.
- observation = 59, 61.

SENSOR = 62.
- auditory = 62, 63.
- visual = 62, 64.

TACTICS = 65.
- defense = 65, 66.
- firing = 65, 67.
- flares = 65, 68.
- offense = 65, 69.
- patrols = 65, 70.

TIME = 71.
- morning = 71, 72.
- afternoon = 71, 73.
- evening = 71, 74.

TRANSPORTATION = 75.
- aerial = 75, 76.
- surface = 75, 77.
  - vehicles = 75, 77, 78.
  - trains = 75, 77, 79.
- water = 75, 80.

WEAPONS = 81.
- artillery piece = 81, 82.
- machine gun = 81, 83.
- mortars = 81, 84.
- small arms = 81, 85.
- tanks = 81, 86.

R REPORTS PILOTS OF HEAVY BOMBERS RETURNING FROM MISSION SIGHTED A 40 – TRUCK CONVOY MOVING SOUTHEAST FROM WAXWEILER (0167 1212443*

REPORTED THAT AN AIR FORCE VISUAL RECONNAISSANCE MISSION THIS AFTERNOON MADE THE FOLLOWING OBSERVATIONS : 1. TRAIN CONSIS 1412448*

•••••••••• 2 INSTANCES OF MISSION

RNACH) AT 1400 HOURS. ABOUT 50 ROUNDS EACH OF 80 MM AND 120 MM MORTAR FIRE WERE RECEIVED DURING THE DAY.

400 HOURS. ABOUT 50 ROUNDS EACH OF 80 MM AND 120 MM MORTAR FIRE WERE RECEIVED DURING THE DAY. 1112444*

•••••••••• 2 INSTANCES OF MM 1112444*

EN ATTENDING NONCOMMISSIONED OFFICER SCHOOLS. C. MORALE OF GERMAN TROOPS IN HIS DIVISION IS GOOD TO VERY GOOD.

•••••••••• 1 INSTANCE OF MORALE 1312443*

N THE AREA UNTIL 1300 HOURS, BUT DID NOT SEE ANY MORE ACTIVITY. THE CURRENT IN THE RIVER WAS SWIFT AND THE WATER WAS LEV

EVIDENCE THAN USUAL DURING THE DAY AND APPEARED MORE CONFIDENT. 2. ABRUPT CHANGE OF ROUTINE OF PERSONNEL ON THE OTHER S 1112441*

SUMMARY STATES : 1. ENEMY OUTPOST PERSONNEL WERE MORE IN EVIDENCE THAN USUAL DURING THE DAY AND APPEARED MORE CONFIDENT. 1512445*

SOUTH IN THE VICINITY OF COORDINATES 2050. 2. 20 MORE SMALL FREIGHT CARS WERE ON SIDING IN THE SAME GENERAL AREA. 3. CON 1512445*

•••••••••• 4 INSTANCES OF MORE 1412448*

NTO LUXEMBOURG WHERE SHE REPORTED TO POLICE THIS MORNING. HER STATEMENT FOLLOWS : I SAW MANY HORSE – DRAWN VEHICLES, PONTOO

S REPORTED IN ALLMUTHUN (040948) DURING THE EARLY MORNING. 3. ENEMY OBSERVATION P(ANES WERE ACTIVE OVER THROUGHOUT THE DI 1412447*

•••••••••• 2 INSTANCES OF MORNING 1512446*

RTED A QUIET DAY. 2. SCATTERED RIFLE, PISTOL AND MORTAR FIRE WERE RECEIVED ALL ALONG THE FRONT. 3. 11 ROUNDS OF LIGHT AR

HOURS. ABOUT 50 ROUNDS EACH OF 80 MM AND 120 MM MORTAR FIRE WERE RECEIVED DURING THE DAY. 1112445*

•••••••••• 2 INSTANCES OF MORTAR 1112444*

LDUP OF ENEMY FORCES ON THE WESTERN SLOPE OF THE MOSELLE VALLEY CONTINUES. THE GROSS DEUTSCHLAND DIVISION HAS AGAIN BEEN

•••••••••• 1 INSTANCE OF MOSELLE 1212445*

THE THINLY HELD AMERICAN LINE. THESE MEN FOR THE MOST PART HAD BEEN ATTENDING NONCOMMISSIONED OFFICER SCHOOLS. C. MORALE

•••••••••• 1 INSTANCE OF MOST 1312443*

Figure 1. Illustration of Key Word in Context (KWIC) Output Listing

and the distinction between the word senses is deemed important, conditional rules are constructed to distinguish between the various word senses. The KWIC is extremely useful in determining what rules will work for this differentiation, because it provides a listing of every unique word along with both the textual context in which the word appears as well as the message identification.

Table 2, below, gives the ENP for the military dictionary in this study. It will be observed that many of the entries are conditional in form in order to differentiate concepts. These conditional entries take the form of IF statements which indicate a search of the text for the specified word contexts forward (+) or backward (-) from the dictionary entry word. Thus, for example, the entry word AIR, in Table 2, indicates that a conditional search is to be made to determine whether the one following word fulfills the stated conditions. The two numbers (1,1) signify that the search begins and ends one word to the right (forward) of AIR. If the word following AIR is FORCE, then AIR is classified as an instance of the concept Army Air Corps. Note that the textual or entry words of the dictionary appear to the left of the colon (:) and the concepts to which they are to be assigned are to the right of the colon (:). For further explanation of the syntax of these entries see the Appendix at the end of this report.

TAGGED OUTPUT LISTING

The initial output of the Inquirer tagging program is the original data plus the categories that have been assigned and stored for future use on some output medium specified by the analyst (e.g., tape, disk, or drum). If the analyst chooses, the output from category assignment may be listed so that the text may be inspected to see how well the category "fits" the data.

Those words which did not receive any categorization are underlined in the listing so that the user knows which characteristics of the data were not handled by any of the dictionary routines. Moreover, after having inspected the listing of the output, the analyst may resubmit the original output for re-tagging by the same (usually updated) dictionary or by additional dictionaries.

Figure 2 provides an example of listed tagged output.

TABULATIONS

Data that have been tagged are tabulated according to specifications provided by the user. The tabulation lists all the categories in which the analyst is interested and provides the raw frequency of occurrence for each concept in the tabulated text. The tabulate program also provides the total number of units in the document such as words, sentences or

Table 2

Entry Name Paragraph (ENP) of the Military Dictionary

```
ACTIVITY          :MØVEMENT.
AFTERNØØN         :TIME.
AGGRESSIVE        :TACTICS.
AIR               :IF WØRD(1,1) = 'FØRCE'
                      THEN ARMY AIR CØRPS
                      ELSE IF WØRD(-1,-1) = 'G2'
                         THEN (ØFFICERS, FRIENDLY)
                         ELSE.
ALMUTHUN          :URBAN.
AMERICAN          :FRIENDLY.
APPEAR            :VISUAL.
AREA              :LØCATIØN.
ARMØR             :BRANCH.
ARMS              :IF WØRD (-1,-1) = 'SMALL' THEN SMALL ARMS ELSE WEAPØNS.
ARMY              :ARMY;
                    IF WØRD (-1,-1) = '15TH'
                       THEN ENEMY
                       ELSE FRIENDLY.
ARRIVE            :MØVEMENT.
ARTILLERY         :ARTILLERY;
                    IF WØRD (1,1) = 'PIECE'
                       THEN (ARTILLERY PIECE; FRIENDLY; EXIT)
                       ELSE IF WØRD (-2,-1) = '299TH'|
                          WØRD (-1,-1) = 'CØRPS'|
                          WØRD (-2,-1) = '28TH'
                          THEN FRIENDLY
                          ELSE ENEMY.
ATTACK            :ØFFENSE.
BANK              :TERRAIN.
BATTALIØN         :BATTALIØN;
                       IF WØRD (1,1) = '295TH'|
                          WØRD (1,1) = '316TH'|
                          WØRD (-1,-1) = 'GUN'|
                          WØRD (1,1) = '915TH'
                          THEN ENEMY
                          ELSE FRIENDLY.
BAULER            :URBAN.
BEFØRE            :TIME.
BEHIND            :LØCATIØN.
BELIEVE           :CØGNITIØN.
BEND              :TERRAIN.
BERG              :URBAN.
BETTINGEN         :URBAN.
BIESDØRF          :URBAN.
BIT               :IF WØRD (-1,-1) = 'A' &
                      WØRD (-2,-2) = 'QUITE'
                      THEN LARGE
                      ELSE.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
BRANDSCHEID          :URBAN.
BRIDGE               :TERRAIN.
BUILDUP              :ØFFENSE.
BULLET               :FIRING.
BURST                :FIRING.
CAR                  :IF WØRD (-1,-1) = 'STAFF'
                         THEN VEHICLES
                         ELSE IF WØRD (-1,-1) = 'FREIGHT'
                            THEN (LØGISTICS; TRAINS)
                            ELSE.
CARLØADS             :TRAINS.
CARRY                :MØVEMENT.
CENTER               :LØCATIØN.
CHANGE               :CHANGE.
CIVILIAN             :CIVILIAN.
CØLLIDE              :MØVEMENT.
CØLØGNE-BØNN-DUEREN  :URBAN.
CØMPANY              :CØMPANY; ENEMY.
CØNFIRM              :CØGNITIØN.
CØNSIDERABLE         :LARGE.
CØNVØY               :VEHICLES.
CØØRDINATES          :CØØRDINATES.
CØRPS                :FRIENDLY: CØRPS.
CP                   :HEADQUARTERS.
CRØSS                :MØVEMENT.
CURRENT              :WATER.
DARK                 :EVENING.
DARKNESS             :EVENING.
DASBURG              :URBAN.
DAWN                 :MØRNING.
DAY                  :MØRNING; AFTERNØØN.
DEFENSE              :DEFENSE.
DESERTER             :DESERTERS.
DETECT               :SENSØR.
DETERMINE            :CØGNITIØN.
DIRECTIØN            :DIRECTIØN.
DISPLAYED            :VISUAL.
DISPØSITIØN          :LØCATIØN.
DIVISIØN             :DIVISIØN.
                      IF WØRD (-1,-1) = 'VØLKSGRENADIER'|
                         WØRD (-1,-1) = 'PANZER'|
                         WØRD (-1,-1) = 'DEUTSCHLAND'|
                         WØRD (-2,-2) = 'ECHELØN'|
                         SØRD (-4,-1) = 'GERMAN'|
                         WØRD (-4,-4) = '320TH'
                         THEN ENEMY
                         ELSE FRIENDLY.
DØUBLETIME           :MØVEMENT.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
DURING              :TIME.
EARLY               :TIME.
EAST                :EAST.
EASTWARD            :EAST.
ECHELØN             :ENEMY; ØRGANIZATIØN.
ECHTERNACH          :URBAN.
ENEMY               :ENEMY.
ENTIRE              :LARGE.
EQUIPMENT           :LØGISTICS.
ESCAPE              :MØVEMENT.
ESCHFELD            :URBAN.
ESTIMATE            :CØGNITIØN.
EVENING             :EVENING.
EVIDENCE            :CØGNITIØN.
EXPLØSIVES          :LØGISTICS.
EXTEND              :EXTENT.
FIRE                :FIRING.
FLARE               :FLARES.
FLEW                :MØVEMENT.
FØLLØW              :IF WØRD (1,1) = 'THE'
                        THEN MØVEMENT
                        ELSE.
FØØT                :IF WØRD (0,1) = 'TRAFFIC'
                        THEN (SURFACE; PERSØNNEL)
                        ELSE.
FØRCE               :ØRGANIZATION;
                     IF WØRD (-1,-1) = 'ENEMY'
                        THEN ENEMY
                        ELSE FRIENDLY.
FØRMATION           :ØRGANIZATION.
FØRMER              :TIME.
FØRTRESS            :DEFENSE.
FRØNT               :LØCATION.
G2                  :INTELLIGENCE; FRIENDLY.
GEICHLINGEN         :URBAN.
GEMUEND             :URBAN.
GENERATØRS          :LØGISTICS.
GERMAN              :ENEMY.
GUARD               :ENLISTED.
GUN                 :IF WØRD (-3,-1) = 'MACHINE'
                        THEN MACHINE GUN
                        ELSE ARTILLERY.
HARASSING           :SMALL.
HEADED              :MØVEMENT.
HEADQUARTERS        :HEADQUARTERS.
HEARD               :AUDITØRY.
HEAVY               :LARGE.
HEINERSCHEID        :URBAN.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
HILL                :TERRAIN.
HØURS               :TIME.
INCREASE            :INCREASE.
INDICATE            :CØGNITIØN.
INFANTRY            :INFANTRY.
INFILTRATION        :ØFFENSE.
JUNCTIØN            :TERRAIN.
KALBØRN             :URBAN.
LARGE               :LARGE.
LAUNCH              :MØVEMENT.
LED                 :MØVEMENT.
LEFT                :MØVEMENT.
LIELER              :URBAN.
LIGHT               :SMALL.
LINE                :LØCATIØN.
LISTENING           :AUDITØRY.
LITTLE              :SMALL.
LØADS               :TRANSPØRTATIØN.
LØCATIØN            :LØCATIØN.
LUXEMBØURG          :URBAN.
MAJOR               :LARGE.
MAN                 :PERSØNNEL.
MANY                :LARGE.
MAP                 :INTELLIGENCE.
MARKED              :LARGE.
MARSHALLING         :LØGISTICS.
MEN                 :PERSØNNEL.
MESSAGE             :INTELLIGENCE.
MILITARY            :MILITARY.
MINES               :DEFENSE.
MINUTES             :TIME.
MISSING             :IF WØRD (-1,-1) = 'RECØNNAISSANCE'
                        THEN EXIT
                        ELSE TACTICS.
MØRNING             :MØRNING.
MØRTAR              :MØRTARS.
MØSELLE             :TERRAIN.
MØST                :LARGE.
MØTØR               :VEHICLES.
MØTØRCYCLES         :VEHICLES.
MØVE                :MØVEMENT.
MØVEMENT            :MØVEMENT.
MUENCHEN-GLADBAD    :URBAN.
NATIØNAL            :IF WØRD (-1,-1) = 'GERMAN'
                        THEN (CIVILIAN; ENEMY)
                        ELSE PERSØNNEL.
NEAR                :LØCATIØN.
NIEDERGECKLER       :URBAN.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
NIEDERSGEGEN        :URBAN.
NIGHT               :EVENING.
NØRTH               :NØRTH.
NØRTHERN            :NØRTH.
NØTES               :CØGNITIØN.
NØW                 :TIME.
ØBSERVE             :VISUAL.
ØBSERVERS           :VISUAL.
ØCCASIONAL          :SMALL.
ØFFENSIVE           :ØFFENSE.
ØFFICER             :IF WØRD (-1,-1) = 'NØNCØMMISSIØNED'
                        THEN ENLISTED
                        ELSE ØFFICERS.
ØLD                 :TIME.
ØPPØSITE            :LØCATIØN.
ØRDER               :TACTICS.
ØRMØNT              :URBAN.
ØUR                 :TERRAIN.
ØVERCØATS           :LØGISTICS.
ØVERHEARD           :AUDITØRY.
PANZER              :ARMØR.
PAST                :TIME.
PATCHES             :LØGISTICS.
PATRØL              :PATRØLS.
PERSØNNEL           :PERSØNNEL.
PILLBØX             :DEFENSE.
PILØTS              :ØFFICERS; FRIENDLY.
PISTØL              :SMALL ARMS.
PLANES              :IF WØRD (-1,-1) = 'ØBSERVATIØN'
                        THEN (ØBSERVATION; FRIENDLY)
                        ELSE ENEMY.
PØINTS              :IF WØRD (1,4) = 'FRØNT'
                        THEN LØCATIØN
                        ELSE CØGNITIØN.
PØLICE              :CIVILIAN.
PØNTØØNS            :WATER.
PØST                :IF WØRD (-1,-1) = 'ØBSERVATIØN'
                        THEN VISUAL
                        ELSE IF WØRD (-1,-1) = "CØMMAND'
                           THEN HEADQUARTERS
                           ELSE IF WØRD (-1,-1) = 'LISTENING'
                              THEN AUDITØRY
                              ELSE LØCATIØN.
PØW                 :PØWS.
PRØJECTILES         :FIRING.
PULL                :MØVEMENT.
QUIET               :SMALL.
RADIØS              :TACTICS.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
RAID                :ØFFENSE.
RAILRØAD            :TRAINS.
RATE                :EXTENT.
REAR                :IF WØRD (1,3) = 'AREA'
                        THEN LØCATIØN
                        ELSE ØRGANIZATIØN.
RECØNNAISSANCE      :RECØNNAISSANCE.
REFLECTØRS          :TACTICS.
REGIMENT            :REGIMENT;
                      IF WØRD (-2,-1) = '295TH'|
                         WØRD (-1,-1) = '320TH'|
                         WØRD (-2,-1) = '352D'|
                         WØRD (-2,-1) = '353D'|
                         WØRD (-2,-1) = '78TH'|
                         WØRD (-1,-1) = '316TH'|
                         WØRD (-1,-1) = '915TH'|
                         WØRD (-1,-1) = '423D'|
                         WØRD (-2,-1) = '942D'|
                         WØRD (-1,-1) = 'THEIR'
                         THEN ENEMY
                         ELSE FRIENDLY.
RELATIVES           :CIVILIAN.
RELIEVING           :TACTICS.
RESERVE             :IF WØRD (-1,-1) = 'IN'
                        THEN TACTICS
                        ELSE LØGISTICS.
RHINE               :TERRAIN.
RIFLE               :SMALL ARMS.
RIVER               :WATER.
RØAD                :TERRAIN.
RØCKET              :ARTILLERY PIECE.
RØSCHEID            :URBAN.
RØTH                :URBAN.
RØTTERDAM           :URBAN.
RØUNDS              :LØGISTICS.
RUINS               :TERRAIN.
RUMØR               :INTELLIGENCE.
RUNDSTEDT           :ØFFICERS.
SAARBRUECKEN        :URBAN.
SALUTING            :MØVEMENT.
SAW                 :VISUAL.
SCATTERED           :SMALL.
SCHEID              :URBAN.
SE                  :LØCATIØN
SEARCHLIGHT         :TACTICS.
SECRET              :INTELLIGENCE.
SECTØR              :LØCATIØN.
SEE                 :VISUAL.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
SEEMS                :CØGNITIØN.
SEVERAL              :LARGE.
SHØRTAGE             :SMALL.
SIDE                 :LØCATIØN.
SIGHTED              :VISUAL.
SIGNIFICANT          :LARGE.
SINGLE               :SMALL.
SLØPE                :TERRAIN.
SMALL                :SMALL.
SØLDIER              :ENLISTED.
SØME                 :SMALL.
SØØN                 :TIME.
SØUND                :AUDITØRY.
SØUTH                :SØUTH.
SØUTHEAST            :SØUTH.
SØUTHWARD            :SØUTH.
SØUTHWEST            :SØUTH.
SS                   :ENEMY.
STRASBØURG           :URBAN.
STRØNGLY             :LARGE.
SUGGEST              :CØGNITIØN.
SUMMARY              :INTELLIGENCE.
SUSPECTS             :CØGNITION.
SYSTEM               :IF WØRD (-1,-1) = 'RAILROAD' THEN TRAINS ELSE.
TACTICS              :TACTICS; ØFFENSE.
TANK                 :ARMØR.
TEAM                 :TEAM.
THINLY               :SMALL.
TIGER                :IF WØRD (-1,1) = 'TANKS'
                         THEN (TANKS; ENEMY)
                         ELSE.
TIME                 :TIME.
TINTESMUEHLE         :URBAN
TØWARDS              :DIRECTIØN.
TØWN                 :URBAN.
TRACKS               :SURFACE.
TRAFFIC              :TRANSPØRTATIØN
TRAIN                :TRAINS.
TRIER                :URBAN.
TRØØPS               :IF WØRD (-4,+0) 'INFANTRY' THEN FRIENDLY ELSE ENEMY.
TRUCKS               :VEHICLES.
UNIFØRMS             :LØGISTICS.
UNIT                 :ØRGANIZATION; ENEMY.
UNLØADING            :MØVEMENT.
UNUSUAL              :CØGNITIØN.
UPCØMING             :TIME.
VALLEY               :TERRAIN.
VEHICLE              :VEHICLES.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
VEHICULAR           :VEHICLES.
VICINITY            :LØCATIØN.
VILLAGE             :URBAN.
VISUAL              :VISUAL.
VØLUME              :EXTENT.
WADED               :MØVEMENT.
WALK                :MØVEMENT.
WAR                 :IF WØRD (-3,-1) = 'PRISØNER'
                        THEN PØWS
                        ELSE.
WATCH               :INTELLIGENCE.
WATER               :WATER.
WAXWEILLER          :URBAN.
WEAPØN              :ARTILLERY.
WEEK                :TIME.
WEST                :WEST.
WESTERN             :WEST.
WHERE               :LØCATIØN.
WITHDRAWAL          :MØVEMENT.
WØMAN               :CIVILIAN.
YARD                :IF WØRD (-1,-1) = 'MARSHALLING'
                        THEN TRANSPØRTATIØN
                        ELSE.
ZWEIBRUECKEN        :URBAN.

0405                :MØRNING.
0450                :MØRNING.
0540                :MØRNING.
0600                :MØRNING.
0630                :MØRNING.
0745                :MØRNING.
0800                :MØRNING.
0830                :MØRNING.
0900                :MØRNING.
0905                :MØRNING.
0910                :MØRNING.
1000                :MØRNING.
1015                :MØRNING.
1019                :MØRNING.
1040                :MØRNING.
1100                :MØRNING.
1115                :MØRNING.
1130                :MØRNING.
1200                :MØRNING.
1300                :AFTERNØØN.
1400                :AFTERNØØN.
1500                :AFTERNØØN.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
1550            :AFTERNØØN.
1600            :AFTERNØØN.
1800            :AFTERNØØN.
1810            :AFTERNØØN.
1830            :AFTERNØØN.
1840            :AFTERNØØN.
1930            :AFTERNØØN.
1940            :AFTERNØØN.
1950            :AFTERNØØN.
2000            :AFTERNØØN.
2015            :EVENING.
2100            :EVENING.
2113            :EVENING.
2130            :EVENING.
2200            :EVENING.
2215            :EVENING.
2245            :EVENING.
2400            :EVENING.
0045            :EVENING.
0300            :EVENING.
0400            :EVENING.
102200          :EVENING.
132300          :EVENING.
132345          :EVENING.
140001          :EVENING.
142400          :EVENING.

908750          :NW1.
995196          :NW1.
040948          :NE2.
064963          :NE2.
0696            :NE2.
0892            :NE2.
9881            :NE2.
996806          :NE2.
821675          :NCW3.
838705          :NCW3.
840673          :NCW3.
847594          :NCW3.
851581          :NCW3.
854584          :NCW3.
871541          :NCW3.
875543          :NCW3.
0167            :NCE4.
1454            :NCE4.
850672          :NCE4.
850673          :NCE4.
854680          :NCE4.
871557          :NCE4.
```

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

873687	:NCE4.
874693	:NCE4.
878698	:NCE4.
880696	:NCE4.
888575	:NCE4.
890570	:NCE4.
8970	:NCE4.
897562	:NCE4.
902692	:NCE4.
930602	:NCE4.
9753	:NCE4.
9241	:SCW5.
9272	:SCW5.
964414	:SCW5.
0450	:SCE6.
1353	:SCE6.
2050	:SCE6.
9052	:SCE6.
915482	:SCE6.
9451	:SCE6.
950440	:SCE6.
952458	:SCE6.
958435	:SCE6.
963436	:SCE6.
965439	:SCE6.
9743	:SCE6.
2129	:SE8.
15TH	:ZØNEN.
V	:ZØNEN.
106TH	:ZØNEN.
18TH	:ZØNEN.
295TH	:ZØNEN.
353D	:ZØNEN.
VIII	:ZØNENC.
26TH	:ZØNENC.
28TH	:ZØNENC.
110TH	:ZØNENC.
112TH	:ZØNENC.
78TH	:ZØNENC.
299TH	:ZØNENC.
352D	:ZØNESC.
9TH	:ZØNESC.
GRØSS	:ZØNESC.
942D	:ZØNESC.
109TH	:ZØNESC.
915TH	:ZØNESC.
116TH	:ZØNES.
212TH	:ZØNES.

Table 2 (Cont'd)

Entry Name Paragraph (ENP) of the Military Dictionary

```
4TH                 :ZØNES.
316TH               :ZØNES.
320TH               :ZØNES.
12TH                :ZØNES.
423D                :ZØNES.
1ST                 :IF WØRD (2,2) = '110TH'
                        THEN ZØNENC
                         ELSE IF WØRD (1,2) = '295TH'
                            THEN ZØNEN
                            ELSE ZØNESC.
2D                  :IF WØRD (2,2) = '295TH'
                        THEN ZØNEN
                        ELSE IF WØRD (1,2) = '316TH'
                           THEN ZØNES
                           ELSE IF WØRD (1,2) = '942D'
                              THEN ZØNESC
                              ELSE ZØNENC.
3D                  :IF WØRD (2,2) = '109TH'
                        THEN ZØNESC
                        ELSE IF WØRD (2,2) = '110TH'
                           THEN ZØNENC
                           ELSE ZØNEN.
RECØRD CØUNT = 00000505, NAME = 0002
CØPY CØMPLETE
```

```
ID FIELD OF LEVEL       2 = 101244

ID FIELD OF LEVEL       1 = 1012441

SENTENCE  1;  2 WORDS
CURRENT ID 1012441

1000        HOURS          .
MORNING     TIME
TIME

SENTENCE  2;  30 WORDS
CURRENT ID 1012441

THE   3D          BATTALION,    109TH        INFANTRY     REGIMENT     REPORTS THAT A  DESERTER  CAPTURED IN  THE
      ZONESC      BATTALION     ZONESC       INFANTRY     REGIMENT                     MILITARY
      ZONE        ORGANIZATION  ZONE         BRANCH       ORGANIZATION                 PERSONNEL
      LOCATION    ORG_STATUS    LOCATION                   ORG_STATUS                   DESERTERS
      AREA_OF_OP  FRIENDLY      AREA_OF_OP                 FRIENDLY

VICINITY  OF  COORDINATES  9241        AT  0540      HOURS  IS  FROM  THE  1ST         BATTALION,    915TH
LOCATION      COORDINATES  SCW5                    MORNING      ZONESC      BATTALION     ZONESC
AREA_OF_OP    LOCATION     COORDINATES  TIME       TIME         ZONE        ORGANIZATION  ZONE
              AREA_OF_OP   LOCATION                            LOCATION    ORG_STATUS    LOCATION
                           AREA_OF_OP                          AREA_OF_OP  ENEMY         AREA_OF_OP

REGIMENT,  352D        VOLKSGRENADIER  DIVISION      .
REGIMENT   ZONESC                      DIVISION
ORGANIZATION ZONE                      ORGANIZATION
ORG_STATUS LOCATION                    ORG_STATUS
ENEMY      AREA_OF_OP                  ENEMY

SENTENCE  3;  15 WORDS
CURRENT ID 1012441

THE   PRISONER  OF  WAR       STATED  HIS  COMPANY       IS  LOCATED  IN  THE  VICINITY  OF  BIESDORF
                    MILITARY              ORGANIZATION                       LOCATION        URBAN
                    PERSONNEL             COMPANY                            AREA_OF_OP      AREA_OF_OP
                    POWS                  ORG_STATUS
                                          ENEMY

9743
SCE6
COORDINATES
LOCATION
AREA_OF_OP
```

Figure 2. Illustration of Tagged output Listing

paragraphs. If the analyst is interested in only a few concepts and not the total number in the dictionary, he may specify which concepts are to be tabulated and which are to be suppressed. In addition, the tabulate program provides index scores which are obtained by the division of the various frequency scores, i.e.,

$$\text{WORD INDEX} = \frac{\text{total assignments of a given concept}}{\text{total words in the entire document}} \times 100$$

ANALYSIS

The various indices, such as the Word Index, provide the basis for determining the "similarity" among the messages in the data base. Several alternative approaches are available for the measurement of similarity and for the determination of the best grouping of messages. In the present study the Pearson product-moment correlation between pairs of messages, computed over the 8 categories of the content analysis, was used as the similarity measure. The eigen vectors of the correlation matrix were then obtained through a principal components analysis, and these vectors were rotated to simple structure using Varimax criteria. Other measures of similarity and other procedures for determining the underlying structure of the messages can be found in Johnson (1967), Anderberg (1972), and Sneath and Sokal (1973).

RESULTS AND DISCUSSION

KEY WORD IN CONTEXT (KWIC) AND FREQUENCY COUNT

As indicated, the first step in the data analysis was to produce a KWIC and a corresponding frequency analysis of the word occurrences in the total set of 40 reports. Figure 1 above displays a sample section of the KWIC for these messages.[1]

The total set of reports contained 3214 word tokens (total words) and 710 word types (a type being defined as any uniquely spelled word form). The frequency distribution is not atypical of that found in most vocabulary counts. Approximately 50% of all of the word occurrences are accounted for by the first 50 most frequent words and there are a large number of single occurring word types. The words HOURS, INFANTRY, ENEMY, COORDINATES and ARTILLERY are among the highest occurring types.

---

[1] The full KWIC and the Frequency Count are available from ARI.

DICTIONARY TAGGING

Figure 2, above, presents a sample of the tagging output produced by the computer routines. The tagged words of this display are from the first report in the message set.[2]

Table 3 summarizes the total number of word occurrences relative to the total words of the report (times a factor of 100) which were tagged in each of the reports for each of the 17 major category divisions of the Concept Name Paragraph (Table 1). These percentages correspond to the Word Index described previously (see discussion of Tabulations under Methods above).

It is the Word Indices for the full 86 categories and subcategories of Table 1 which will be submitted to analysis (see Table 4 for the Sequence Numbering used).

REPORT FACTOR STRUCTURE

Table 5 summarizes the factor structure obtained from a principal components analysis of the report correlations based on the full set of 86 categories of the dictionary. (Three categories have zero occurrences of identified words. These categories would have yielded zero divisor checks in calculating the factor scores, and for this reason they were dropped.)

Five factors accounting for approximately 85% of the total report variance had eigen values greater than 1.00 and they were rotated by Varimax criterion to simple structure.

Factor I identifies reports of large scale enemy troop movements or locations of considerable strategic importance. Factor II identifies reports of vehicular and small scale movement all along the forward edge of the battle area. Factor III identifies those reports dealing with unusual small arms fire. Factor IV identifies deserter and POW reports of lesser reliability and Factor V identifies civilian, prisoner of war interrogation team and reconnaissance reports from reliable sources.

Within each factor, the factor coefficients for each intelligence message indicate the relative weight of the message on the factor. For example, Factor II loadings (coefficients) diminish when the report content deals with foot or patrol activity as compared to convoy and logistic or tactical support traffic.

An inspection of Table 5 indicates that the dictionary successfully organized the message sample into a set of factor groupings which are logically coherent despite the relatively small size of the test dictionary employed. Each of the factors of this report structure represents independent dimensions of classification of the total message sample with successive dimensions accounting for decreasing amounts of the report similarities.

---

[2]The complete tagging outputs for all reports are available from ARI.

Table 5

Summaries of Major Category Tags for Word Index

MESSAGE NUMBER	DIRECTION	TIME	LOGISTICS	MOVEMENT	PLANES	TRANSPORTATION	ORGANIZATION	ORGANIZATIONAL-STATUS	SENSOR	EXTENT	CHANGE	WEAPONS	TACTICS	PERSONNEL	BRANCH	INTELLIGENCE	AREA OF OPERATIONS
1		8.51					12.76	12.76						4.25	2.12		23.40
2	2.38	9.52		2.38		4.76	4.76	9.52	2.38				2.38	2.38	2.38		16.66
3		17.85				7.14	7.14	7.14	3.57						3.57		17.85
4	.96	7.69		4.80		7.69	2.88	5.76	3.84				3.84	.96	.96	1.92	10.57
5		4.65					4.65	4.65		2.32			2.32	6.97	2.32		18.60
6	2.73	2.73	1.36	2.73			5.47	2.73		1.36			2.73	2.73	1.36	1.36	13.69
7	5.40	16.21	2.70				2.70	5.40				8.10	5.40		2.70		8.10
8		6.81	2.27	2.27			4.54	9.09		6.81		6.81	4.54		4.54	2.27	6.81
9	3.33	13.33					6.66	10.00	6.66			3.33		3.33	6.66		26.66
10	4.44	13.33		6.66		4.44	2.22	6.66	4.44	2.22		2.22			6.66		11.11
11	3.33	13.33		3.33	3.33	3.33	3.33	16.66	3.33	3.33			3.33	6.66		3.33	16.66
12		22.91		2.08		2.08	2.08	6.25	4.16	6.25		2.08	4.16		2.08		12.50
13	1.40	2.81		1.40		1.40	11.26	14.08					2.81	1.40	1.40	4.22	14.08
14		12.12					9.09	15.15	3.03				9.09	3.03	6.06		9.09
15		15.38		3.84			7.69	7.69	3.84	3.84			3.84		3.84		19.23
16		4.20					9.24	10.92		1.68			2.52	3.36	.84	.84	9.24
17		20.33				8.47	1.69	5.08	3.38						3.38		18.64
18		16.66		3.33		1.66	1.66	5.00	1.66	3.33	1.66	3.33	5.00		1.66	5.00	15.00
19	3.17	7.93		4.76			14.28	11.11						1.58	3.17	7.93	11.11
20		20.00				4.00	4.00	4.00	16.00					4.00	4.00		16.00
21		9.75		2.43			9.73	7.31	2.43				2.43	2.43	2.43		26.82
22	3.33	26.66		3.33		3.33	3.33	3.33	3.33	3.33					3.33		13.33

Table 3 (Cont'd)

Summaries of Major Category Tags for Word Index

MESSAGE NUMBER	DIRECTION	TIME	LOGISTICS	MOVEMENT	PLANES	TRANSPORTATION	ORGANIZATION	ORGANIZATIONAL-STATUS	SENSOR	EXTENT	CHANGE	WEAPONS	TACTICS	PERSONNEL	BRANCH	INTELLIGENCE	AREA OF OPERATIONS
23	1.13	6.81		1.13		1.13	22.27	3.40		1.13		1.13	1.13	1.13	1.13		25.00
24		17.39		4.34		4.34	4.34	3.34	4.34						4.34		30.43
25		11.11					8.88	11.11	2.22	2.22			4.44		6.66	2.22	17.77
26	1.76	3.53	2.65	4.42		3.53	1.76	2.65	2.65	4.42				4.42	.88		8.84
27	1.63	11.91	4.91	3.27		6.55	4.91	6.55	1.63	4.91				1.63		3.27	13.11
28	2.77	11.11		5.55		5.55	2.77	2.77	8.33	2.77				2.77	2.77		13.88
29		8.62					6.89	8.62					3.44	3.44	3.44	3.44	13.79
30	1.21	14.63		1.21						7.31	1.21	3.65	4.87		1.21	2.43	6.09
31		4.80					12.50	13.46		.96			.96	.96	2.88	1.92	12.50
32		10.38	1.29				2.59	10.36	1.29			1.29	6.49		5.19	2.59	9.09
33		6.75		2.70		2.70	8.10	10.81	5.40	4.05	1.35		2.70	2.70	1.35	2.70	22.97
34		18.51				7.40	7.40	7.40		3.70					7.40	3.70	11.11
35		4.54	3.03	1.51		6.06	3.03	6.06						4.54	1.51		7.57
36		25.00					3.12	3.12		12.50	3.12	3.12	3.12		3.12	6.25	9.37
37		10.58	1.17	3.52		2.35	1.17	5.88	3.52		1.17			3.52	1.17	1.17	10.58
38		10.16		1.69			8.47	8.47		1.69		1.69	3.38	1.69	5.08	1.69	6.77
39		8.33		2.08		2.08	2.08	6.25		2.08	2.08			4.16		6.25	6.25
40		9.09	.84	1.94	.84	.84	5.19	8.44	1.29	2.59			4.54	1.94	.84	.84	9.09

Table 4

Concept Categories by Sequence Number

1. North West
2. North East
3. North Central West
4. North Central East
5. South Central West
6. South Central East
7. South West
8. South East
9. Zone North
10. Zone North Central
11. Zone South Central
12. Zone South
13. Coordinates
14. Zone
15. Deserters
16. Enlisted
17. Officers
18. POWs
19. Vehicles
20. Trains
21. Decrease
22. Increase
23. North
24. East
25. South
26. West
27. Large
28. Small
29. Cognition
30. Reconnaissance
31. Army
32. Army Air Corps
33. Corps
34. Division
35. Regiment
36. Battalion
37. Company
38. Headquarters
39. Team
40. Enemy
41. Friendly
42. Civilian
43. Military
44. Location
45. Terrain
46. Urban
47. Armor
48. Bombers
49. Observation
50. Auditory
51. Visual
52. Defense
53. Firing
54. Flares
55. Offense
56. Patrols
57. Morning
58. Afternoon
59. Evening
60. Aerial
61. Surface
62. Artillery
63. Infantry
64. Water
65. Artillery Piece
66. Machine Gun
67. Mortars
68. Small arms
69. Tanks
70. Direction
71. Time
72. Logistics
73. Movement
74. Planes
75. Transportation
76. Organization
77. Organization Status
78. Sensor
79. Extent
80. Change
81. Weapons
82. Tactics
83. Personnel
84. Branch
85. Intelligence
86. Area of Operations

Table 5

Report Structure Using Concept Category Tags

FACTOR I

REPORT	LOADING	NOTES
31	.85	Move of 352d Artillery Regiment Headquarters
13	.84	Buildup of large enemy forces on western slope of Moselle Valley
16	.83	320th Regiment in reserve. 212, 316 & 423 Regiments calling up reserves
19	.78	Southward Movement of the 116th Panzer Division
14	.74	Enemy patrol in vicinity of 229th Artillery Battalion Command Post
38	.73	Panzer and infantry divisions in front of 4th Infantry Division to launch major offensive
29	.66	Enemy map from 26th Volksgrenadier Division found
32	.65	Enemy keeping artillery well back from front and no counter fire during past week
1	.64	352d Volksgrenadier Division located near Biesdorf
25	.63	*Large formation of enemy infantry around Gemuend preparing for raid into 110th sector
40	.62	Considerable vehicular activity all along front, enemy observation planes, patrols and searchlights reported

FACTOR II

REPORT	LOADING	NOTES
20	.82	Truck movement at coordinates 878698
3	.80	Horse-drawn vehicles and staff car between coordinates 890570 and 897562
22	.75	Vehicular movements vicinity of coordinates 854680
10	.75	Horse-drawn vehicles vicinity of 888575 and 930602
28	.74	Vehicles vicinity of 880696
17	.71	Tracked vehicles, trucks, motors, possibly tanks between coordinates 874693 and 9272
24	.69	Vehicle movement at coordinates 878698 and vicinity 873687
37	.62	Foot traffic vicinity of 965439, 963436 and 958435
2	.58	Enemy patrol movement in vicinity of 851581

*Fictitious report

Table 5 (Cont'd)

Report Structure Using Concept Category Tags

FACTOR III

REPORT	LOADING	NOTES
30	.92	Unusual, intermittent, harassing small arms fire along 112 sector front.
36	.84	Marked increase of harassing small arms fire along 109, 110, 112th Regiment fronts
12	.73	Rifle, pistol and flare activity along 112th sector front
7	.72	Enemy rifle, pistol and mortar fire along northern sector of 4th Division
8	.65	Scattered rifle, pistol mortar, and light artillery fire along 28th Division front
18	.64	Indiscriminate rifle and pistol firing, unusual motor activity and aggressive enemy patrol activity along 106th Division front

FACTOR IV

REPORT	LOADING	NOTES
23	.78	*Enemy deserter reports machine gun emplacement moved to vicinity 850672 protecting bridge approach
6	.73	*Enemy deserter reports movement of 15th Army to Cologne-Bonn-Dueren area. Von Rundstedt orders withdrawal
5	.72	*POW team reports reliable civilians state that mines laid vicinity of Bauler and enemy soldiers in Niedergeckler
21	.68	Wounded POW reports Headquarters of 26th Division moved up to Eschfeld-Roscheid area
33	.59	POW reports 1st Battalion, 259th Regiment in Ormont area. Listening post reports increase in vehicle movement

FACTOR V

REPORT	LOADING	NOTES
26	.71	POW team reports highly reliable civilian reports river crossing equipment SS Troops, artillery and vehicles moving West from Bitburg
35	.71	*POW team reports German national reports marshalling yards at Zweibruecken tied up for six hours
27	.50	Air Force reconnaissance reports considerable activity in marshalling yards at Trier
39	.40	Abrupt change of routine on other side of Sauer River suggests arrival of new troops

*Fictitious reports

In order to make clearer the origins of these factors, examine the character of the two highest loading reports of Factor I: Reports 31 and 13. From Table 3 note that Report 31 contains word occurrences which pertain to nine of the 17 major categories of the dictionary, and that Report 13 contains word occurrences pertaining to 11 of the categories. The two reports had word occurrences pertaining to eight common categories: TIME, ORGANIZATION, ORGANIZATIONAL STATUS, TACTICS, PERSONNEL, BRANCH, INTELLIGENCE, and AREA OF OPERATIONS. On the basis of the number of shared categories indicated by their word entries, the two reports have a correlation of .91, indicating a considerable similarity of content.

Whatever structure emerges is obviously dependent on the adequacy of the dictionary which is used to process the reports. The factor groupings of the intelligence reports are based solely on the overlaps of the concept categories which occur among the reports. If the dictionary categories as defined are not relevant, the ultimate structure which emerges will be useless. Additionally, the interpretation which is to be given to the structure must be made in terms of the categories which are employed. Thus, for example, the reports grouped under FACTOR II significantly share content occurrences in the dictionary category of Transportation. Factor III reports most prominently share the content of categories Extent, Change, and Weapons. Such a basis for classification is eminently reasonable. Shared content is a practical means of establishing report groupings of significance for intelligence retrieval and analysis provided that the content is meaningfully defined. As indicated, however, the procedure allows an analyst to redefine and change the categories of the dictionary as he chooses and as required by the retrieval or organizational needs of the intelligence situation for which the method is to be used.

Factor I is the most important factor in terms of amount of report similarity for which it accounts. It clearly identifies tactical intelligence of considerable significance. It successfully sifts out those reports of lesser tactical importance while nonetheless including reports not having obvious tactical implications unless seen in the context of the other reports of this factor. Thus, for example, the reports of a small enemy patrol discovered near the 229th Field Artillery Battalion command post takes on tactical significance when included in the taxon including reports of large-scale troop movements and enemy build-ups. Similarly, the otherwise negative report of an absence of enemy artillery fire, especially counter-battery fire, when seen in the context of this factor, implies an enemy stratagem, although otherwise it could imply (as was actually assumed in 1944) an enemy ammunition shortage and defensive posture.

Factors I, II, and III in decreasing order of importance identify various kinds of enemy activity from troop movements and immediate attack threats through truck and convoy movements to small arms activity. Factors IV and V pull together hearsay reports of varying degrees of reliability. Three of the seven false reports (Reports 6, 21, and 23) added

to the message set are found with high loading on Factor IV. Of the remaining four fictitious reports, one (Report 25) is found with lower loading on Factor I and another (Report 35) on Factor V. False reports 9 and 15 did not appear with significant loadings under any of the five factors. It is reasonable to expect that deserter, civilian, and POW reports would be of lesser reliability than reconnaissance or interrogation reports, and it is on the basis of these report sources that the computer algorithm has grouped the reports.

CONCLUSIONS AND RECOMMENDATIONS

The application of a set of procedures based on the content analysis of a tactical intelligence message set led to the identification of a multidimensional message structure. This logically coherent structure could provide assistance to intelligence analysts in the organization and analysis of the data in the message set. However, the content-analytic procedures must be refined.

One clear inadequacy of the present dictionary definition language, as used in producing the above results, is the current inability to define number ranges. Thus, all clock times, map coordinates, and other numbers must be represented as uniquely occurring forms. Additionally, the entire General Inquirer System is programmed only for the IBM S/360 or S/370 computer. Apart from system problems, the generality of the present results are limited by the relatively small data base employed and by the use of a specialized dictionary based on and adapted to that data base. Nevertheless, we believe the results indicate sufficient promise for this intelligence data organizational scheme to warrant further investigation.

References

Anderberg, Michael R. Cluster analysis for applications. Air Force Systems Command, Kirkland AFB, 1972. (AD 738 301).

Johnson, S. C. Hierarchical clustering schemes. Psychometrica, 1967, 32, 241-254.

Miller, J. P., and Psathas, G. A 1401-1311 tagging problem for the General Inquirer. Behavioral Science, 1968, 13, 166-167.

Miron, M. R., Patten, S. M., and Halpin, S. M. Determination of the Structure of Combat Intelligence Ratings. U.S. Army Research Institute for the Behavioral and Social Sciences, Technical Paper 286, 1977.

Patten, S. M. An inductive taxonomy of combat intelligence data. U.S. Army Research Institute for the Behavioral and Social Sciences, Research Memorandum 74-14, December 1974.

Sneath, Peter H. and Sokal, Robert R. Numerical Taxonomy. San Francisco: W. H. Freeman & Co., 1973.

Stone, P. J., Bales, R. F., Namenwirth, J. Z. and Ogilvie, D. M. The General Inquirer: A Computer System for Content Analysis and Retrieval Based on the Sentence as a Unit of Information. Behavioral Sciences, 1962, 7, 484-498.

Stone, P. J., Dunphy, D. C., and Bernstein, A. The Analysis of Product Image. In P. J. Stone et al. (Eds.), The General Inquirer: A Computer Approach to Content Analysis. Cambridge, Mass: MIT Press, 1966, 619-627.

Stone, P. J., Dunphy, D. C., Smith, M. S., and Ogilvie, D. M., The General Inquirer: A Computer Approach to Content Analysis. Cambridge, Mass. MIT Press, 1966.

APPENDIX

SPECIFICATIONS FOR THE ADAPTATION OF THE GENERAL INQUIRER SYSTEM TO ARI ADP EQUIPMENT

Introduction

The methodology for Data Organization discussed in this report is an adaptation of the General Inquirer System which uses a Content Analysis Dictionary to classify words in a document according to concept names, or categories, as defined by the user. Building a new dictionary for documents in a particular environment represents significant effort and time but it allows an individual analyst to investigate his own theories by modifying existing files. The proposed update routine should expedite any changes or variations to existing dictionaries and is easily adapted to program interactive terminal use.

The suggested system consists of several programs or modules which can be executed singly or in sequence as specified by input control cards or keyboard type-ins. In such a system, the minimum configuration would include:

1. Text Editor/KWIC
2. Dictionary Compiler
3. Dictionary Update
4. Tagging Run
5. Retrieval Edits
6. Statistical Options

The system flow and interconnections are shown in Figure A-1. The individual function blocks are described below.

Text Editor/KWIC

This program scans the original documents as entered at a remote terminal, checks level identifiers, punctuation, flags numerical strings, and formats data for writing on tape (or disc). Any serious errors are to be displayed and identified by document ID and sentence number.

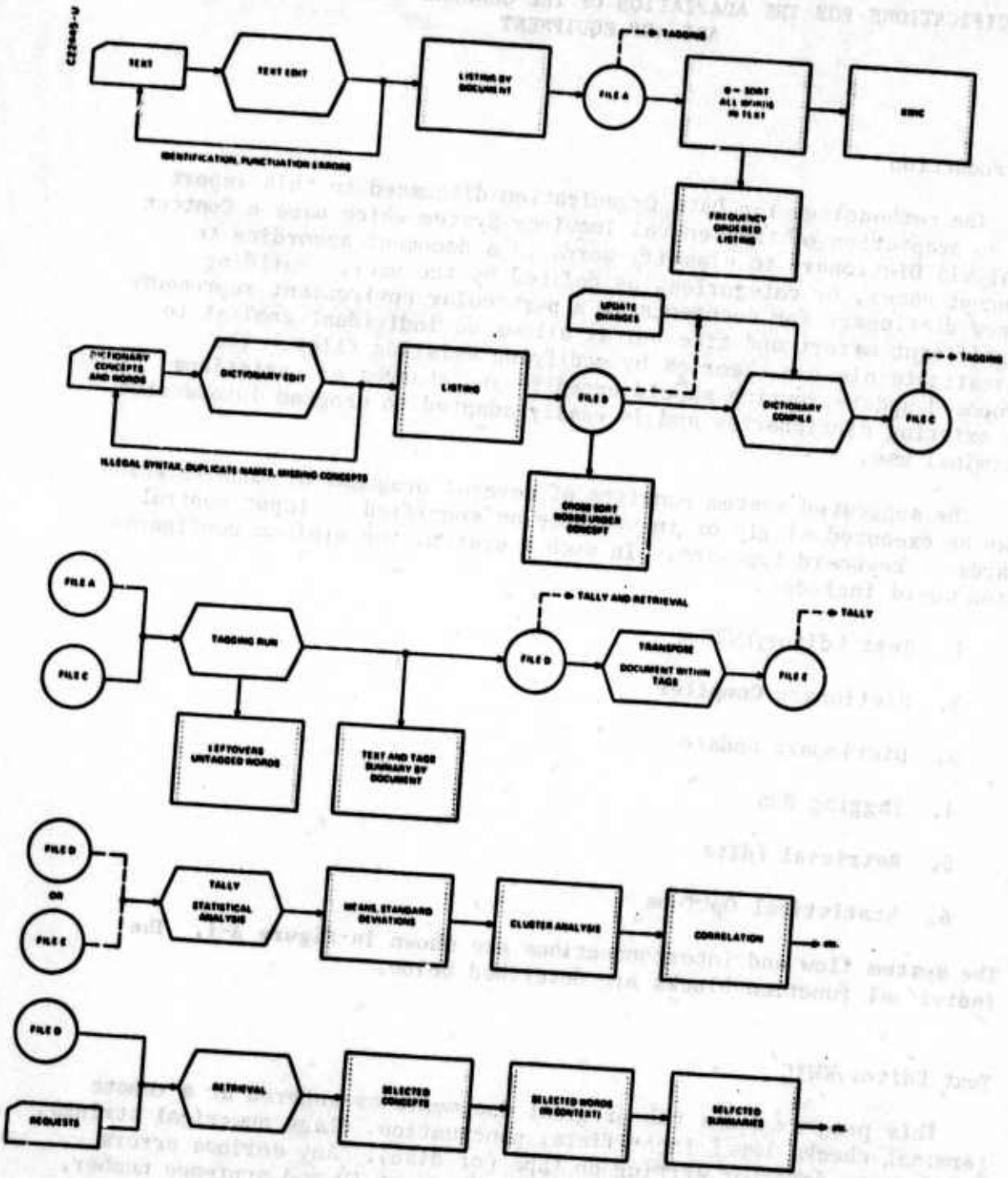

Figure A-1. System Flow for Adaptation of General Inquirer

The program makes a word count as the input is processed and prints a frequency ordered listing followed by an optional KWIC (Key Word in Context) in alphabetical order, along with the document identification field. The sorting above involves first the frequency count and second the alphabetized appearance of each word found in the text.

To shorten the output, certain restrictions may also be imposed to cause the program to ignore common prepositions, articles and auxiliary verbs, since the present purpose is not to analyze syntax.

Dictionary Compiler

The dictionary input (initially on cards) consists of three sections in the following order:

Concept Names as defined by the user, with their organization and structure determined by their starting position on the card. Each name must be unique. Multiple-word names connected by hyphens are permissible.

User tag definitions for precoded categories.

Word definitions which specify the concept names to be applied and under what conditions.

The syntax rules for word definitions (described later) can be fairly complex so this program should be run separately to ensure that there are no format errors, duplicate concept names or incomplete word statements. If no serious errors are found, the compiled dictionary is written on tape with an appropriate ID for each of the three sections mentioned above.

Dictionary Update/Compile

This program provides for modifications to be made to an existing dictionary already on tape. The general types of modification are the usual DELETE, ADD, and REPLACE functions with the capability of rearranging the concept name structure by combining names under a new category, moving a name from one level to another, creating new names, or indicating one of more names as synonymous. These functions eliminate the necessity for making many changes in the concept names as specified in the word descriptions.

After all changes have been made, the updated version is passed to the compiler for a new run.

Dictionary Update

After an initial tagging run, it may become apparent that some concept names have low frequency counts and should be combined, or that other categories are not structured properly. To avoid making changes to the original card deck for the dictionary concepts and word definitions we envision a program to add, delete, or replace entries. The input may be punched on cards or optionally entered at a terminal keyboard.

Since the concept structure must be flexible, especially in the early stages of data organization, some possible functions contemplated are:

NAME	A, B, ...	Define new category names.
ERASE	A, B, ...	Delete names from the table, including any lower level structures.
RENAME	A B.	Redefine A to be B (new name). A is deleted.
EQUIV	A B.	Equivalence A = B.
MOVE	A B.	Move A under B. If no B is specified, move means to assign to major level.
BREAK	A B, C,.....	Split A into new subcategories B, C, etc.
MERGE	A B, C,.....	Merge B, C, .. under category A. B and C are erased. If B or C = A then combine at level A. If no B or C then merge all sublevels.

Care must be exercised in changing the actual structure of concept names. If a name is erased, the corresponding word definitions may no longer be valid and require modification. Such modifications are also possible, with certain restrictions in procedural order. Deletions or replacements should be made first, then the additions.

Since the modifications above are being made to a previously compiled (error free) dictionary, stringent program checks are needed to detect inconsistencies before the updated version is recompiled and used.

Tagging Run

Input to the tagging program consists of edited text and a compiled dictionary. The most efficient method of processing is to store the entire dictionary in core memory and process the text, one report or document at a time. If core memory is limited, the run time will be considerably longer and other techniques will have to be evaluated.

Each word in the text is first matched against the words defined in the dictionary. If no match is found, the word is added to the "left over"* list and the next word is processed. If a match is found, the concept name(s) associated with the definition are appended to the text word and will appear in the output listing.

Since the summaries or counts are kept for each sentence within a report or document, the level identification and sentence number are sufficient to identify each appearance of the tagged words. The total number of sentences per report or document and the number of words per sentence are also retained. Relative summary counts can be calculated based on either of these totals.

Output from this program is a listing for each document, each line of text with one or more concept names at all levels appearing directly below the tagged words.

A summary for each document appears at the end with raw and relative frequencies given for each category in low to high level organization. These values are also written on tape with the appropriate identifiers and can be used by other modules such as the Transpose and Statistical Analysis programs.

On the first tagging run with a new dictionary, this format provides a complete reference for checking the concept structure and finding possible redundant definitions in the word specifications.

Statistical Options

From the experience gained in the test project, we recommend that summaries of the total number of words tagged in a report or document relative to the number of words in the document, i.e., the word index of the Method section, be used as the basic numerical index for correlational analysis of the report patterns. These indices may be employed in any standard factor analysis routine, a FORTRAN version being appended to the tag file output of the content analysis routines.

---

*Untagged.

Retrieval Edits

Although retrieval may not be specifically required, it is desirable in order to complete the system. The proposed system would use a request format similar to the dictionary syntax except that a different set of action words to describe the various types of output desired would be employed. These control words are LIST, CONCEPTS, and TAB. Each request is treated separately. The previously tagged text is scanned and a report is constructed according to the conditions satisfied.

LIST	All sentences.
CONCEPTS	Words tagged.
TAB(n)	Summary tables for level n.

The form for a request may be compound or conditional as in the word definitions (minus the "word:"). The program prints the sentences, concept names and tagged words, or summary tables for the conditions specified.

Example: If TAG (n_1,n_2) = OFFICER

THEN TAB (2)

ELSE.

In this context n_1 and n_2 refer to the word count relative to the beginning of the sentence.

Input Conventions

A document can be defined as a unit of text containing one or more sentences, such as a message (report), abstract, or paragraph. Analysis is done for each sentence so the grouping can be arbitrary. Although input is basically free form, certain punctuation marks are to be reserved for special functions such as identification, comments, or user tags.

Terminal Punctuation [. ! ?]

The end of a phrase or sentence is defined by a period, exclamation mark, or a question mark. Normal rules of punctuation are followed except when ending a quotation such as . . .".

Special Delimiters [$ [ ] (( ))(} * /]

$---$ Characters between dollar signs are considered to be document titles. They are used as output headers and are not tagged.

[---] Identification codes are enclosed between left and right brackets and, if used, must appear at the beginning of a document. The level of identification is indicated by the number of brackets such as [[PZ123]] is the ID at Level 2.

((---)) Double sets of parentheses can be used for hand coding a synonym for the preceding word, which is not defined in the dictionary. An example might be:

.....TIGERS ((TANKS))

where the word TANKS does not appear in the original text, but is described in the dictionary. The word TIGERS will be tagged as if it were equivalent to TANKS.

{---} Braces are used to set off comments or explanations written by the coder and are not tagged. Since these characters do not appear on a keypunch the convention could be #(and #) to produce the left and right braces. For teletype or keyboard characters the equivalents might be Δ and Δ .

***** An asterisk is the signal for end-of-document at level one. Corresponding to the identification of levels, the number of asterisks define the level index. At least one space must precede each set of asterisks in ascending order.

Example: ..... the end. * **

The single * flags the end of level 1 and ** is the end of level 2, where the level index, low to high, goes from 1 to n - 10. This hierarchy may seem to converse of formal ordering but makes it convenient to add a new level without reordering identification of the "lower" level documents.

/ A slash followed by one or more letters is inserted by the coder to hand tag special words or numbers. If a space immediately follows the / the word itself is not tagged and may be followed by a word inserted by the coder.

Normal Punctuation [( ) : ; , " ']

Single sets of parentheses are treated as enclosing a conventional parenthetical phrase. The other characters are also handled in the usual sense and quote marks must appear in pairs. All of the above also act as word delimiters (with or without surrounding spaces), except for the apostrophe which is considered as part of the word, e.g., "o'clock" or "don't."

Miscellaneous [- . .. ...]

The hyphen has two possible uses. If used to separate compound words such as HALF-BAKED, there must be no spaces and the word is treated as a single unit. To indicate a pause or break, one or more hyphens may be used if surrounded by blanks.

A single period used as a decimal point can be recognized as part of a decimal number with certain restrictions. 3.1416 or 3.0 will be treated as a number string, but 10. will appear as an end of sentence marker.

The convention to indicate abbreviations is flexible, but the simplest method, which we recommend, is that the periods be ommitted so as not to be confused with sentence terminal punctuation.

Card Format

Each document should start on a new card in order to facilitate listings, corrections, or the insertion of a new level of identification. The proposed scan program will not require it, however.

The original text is punched as if for typing, with two exceptions.

- The character set is assumed to contain only upper case characters, the digits 0 through 9, and standard punctuation marks or characters belonging to the ASCII code* (limited EBCDIC).**

- Continuation cards are considered as an extension (past column 80) of the preceding card. Words should be hyphenated from line to line only if the hyphen is part of the word. A word ending in column 80 should be followed by a card with a blank in column 1.

One or more blanks, or any of the normal punctuation marks are word delimiters. As mentioned above, the exceptions are hyphens within word strings, a single period in a number string, and the apostrophe.

---

* American Standard Code for Information Interchange - a 7 or 8 bit code for teletypes/keyboards.

** Extended Binary-Coded-Decimal Interchange Code (8-Bit Code).

Special Cases

[= & % @ =] are legal, but if not set off by spaces will be considered as part of a word.

[\↑↓] are characters which might be used for special controls if data are entered on-line via a teletypewriter or keyboard.

Illegal codes [¢ ¬ ⌊ |] will be ignored but in any event are normally not available on teletypewriters and most terminal keyboards.

Content Analysis Dictionary

Specifications for the dictionary consist of three separate sections - Concept Names, User Tags, and Word Definitions. The first step in compiling a new dictionary would be to run a KWIC on part or all of the text and consider the general groups of interest, the high frequency words, and any other phrases or words which require special consideration because of their unusual or local context. In general, articles, prepositions, and auxiliary verbs are of little interest and would be specified as part of a NOT table.

Concept Names

The proposed adaptation would define the correspondence between concept names and tag numbers by the expression:

NAME = Tag Numbers.

This relationship can be a simple one-to-one correspondence or a more complicated one to describe major, minor, and one or more subcategories in the form:

Major name = t_1

Minor name = t_1, t_2.

Subdivision = t_1, t_2, t_3.

Etc.

A concept name must start with a letter and contain only the letters A - Z, digits 0 - 9, and the special character "hyphen." Each name must be unique and should contain no more than 20 characters, and have a reasonable limit, such as 10, on the levels of categories.

User Tags

In some cases, certain words or numbers in the text cannot be easily described by word definitions and yet the human editor or coder wants to mark these as belonging to a specific category. An example from the messages might be to tag all map coordinates with a C, such as:

".....found in the vicinity of 854584/C."

To equate the C to the desired concept name, already appearing in the level organization under Concept Names, Use:

C = COORDINATES.

Hand coded tags must be single letters but more than one may be applied to a single word such as:

.....vicinity of 854584/CR.

where C is as above and R might be assigned to indicate an area east of the river such as:

R = RIVER-EAST.

If no letter immediately follows a "word/" the tagging of that word will be inhibited. Such instances are usually followed by a coder's insert as in the following example:

..... of automatic/ ((pistol)) or machine gun fire.

The word pistol will be tagged according to the category assignment in the dictionary but the word "automatic" will not.

Dictionary Words

Word definitions or statements comprise the major part of the dictionary. They do not contain the conventional synonyms, uses and explanations but are declarations which specify to which category the word belongs, and the rules or conditions for classification. The general form is:

Word: Operation String.

Since words in the text are subject to "chopping", common plural forms for nouns and tense forms for regular verbs need not be entered. The algorithm used will be described later.

The operation string may be simple, compound or complex. The simplest form is:

Word: Concept Name.

Example: SNOW: WEATHER.

A compound expression is a series of Concept Names separated by semi-colons (;) and ending with a period (.).

Example: SNOW: VISIBILITY; TRAFFICABILITY; PRECIPITATION.

A complex string uses key words to describe conditional tests for classification assignment in the form:

Word: IF Conditional expression

THEN Clause A.

ELSE Clause B.

The interpretation is as follows: IF the conditions in the expression are true (i.e., match equally or produce a logically true result), THEN process Clause A; ELSE (otherwise) process Clause B. Clause A or Clause B may, in turn be another conditional expression. The directives IF, THEN and ELSE are reserved for use as control words and may not appear as Concept Names.

Another key word is EXIT which signals the logical end of a conditional path as distinguished from the period at the end of the operation string. If Clause A does not end with the EXIT directive the path drops through to the end of the ELSE branch, unless it is a nested conditional expression.

The control word CALL allows an equivalent definition to be used rather than having to write the same statements several times. Examples are synonyms or abbreviations.

Example: RGT: CALL REGIMENT.

Options possible under the IF expression are TAG, WORD, CHOP, and two special ones, ID and NUMB. The form allows for equal, not equal, and combinations of the logical functions, AND, OR, NOT.

Example: IF key (n_1, n_2) = a op

key (n_1, n_2) = b

THEN . . .

where key is one of the options mentioned above; n_1 and n_2 define the search range in the text relative to the current word. n_1 or n_2 refer to word counts and the search can be backward (negative) or forward (positive). If n_1 and n_2 are not specified the entire sentence is searched.

"Op" refers to one of the logical tests where the symbols used are:

for NOT

& for AND

/ for OR

The tests are performed in the order listed.

Unless the text has been previously tagged, "IF TAG" in a forward direction should be used with care since ambiguity is possible in the case of complex expressions.

```
Example:  PATROL:  IF WORD (-3,-1) = ENEMY /
                       WORD (-2,-1) = GERMAN
                   THEN ENEMY-UNIT: EXIT
                   ELSE IF TAG = SECTORA
                   THEN RECON
                   ELSE.
```

When the word PATROL or PATROLS is encountered in the text, a search is made in the preceding three words for a match with ENEMY, or in the preceding two words for a match with GERMAN. If either word is found, then assign the concept name ENEMY-UNIT and exit, i.e., end of definition.

If neither match is made, then search the entire sentence for concept SECTORA. If a match is found, assign the tag RECON since SECTORA is known to refer to friendly territory.

"CHOP" is similar to WORD except that the match word determines the mask or number of characters to be used in the search.

```
Example:  DIRECTION:  IF CHOP (-100, -1) = NORTH
```

When the word DIRECTION is encountered, a search of the preceding words is made, beginning at the first word in the sentence (arbitrarily specified by the -100 range limiter) and will pick up any of the following: NORTH, NORTHEAST, NORTHWEST, NORTHERLY, NORTHERN, etc.

The other key word "ID" has a slightly different interpretation. The parameters n_1 and n_2 specify the first and last characters inclusive, to be checked in the current ID field. An optional third parameter n_3 refers to the level of identification, with a default value of 1.

Example: HOURS: IF ID (1,2,2) = 10

THEN DEC10

ELSE.

The first two characters in the current ID field, level 2, are compared with 10. If a match is found, assign the name DEC10, otherwise do nothing.

Chopping Algorithm

To reduce the size of the dictionary (and the work involved in its preparation), the stragegy used in the General Inquirer is to define an algorithm which will "chop" a text word by removing the most common suffixes to find the root. The corresponding rules for prefixes is much more difficult and will not be attempted.

During the tagging run, each word in the text ($\leq$ 20 characters) is first matched against the words defined in the dictionary. If an exact match is found, the tag numbers are affixed to the word and saved for printing purposes.

Ideally, the index for any match on the first four letters could be saved to eliminate a complete re-scan in the second search.

The next step is to subject the word to a series of tests for the most common endings, double letters, and adverbial endings by removing the letters 's', 'ing', and 'ed'. If none of these truncations is possible the word is considered to be undefined and no tagging is done. If chopping is successful, the shortened word is rematched against the dictionary.

The obvious exceptions are the non-standard forms of irregular verbs such as come/came and some noun plurals such as man/men.

Problem Areas

The General Inquirer has one serious shortcoming - the inability to recognize or perform tests on numerical strings. One solution is to provide for automatic tagging of "words" beginning with any digit 0 through 9. This process would not conflict with user's tags or any qualifiers in the dictionary word definitions.

Internal tags would be generated for the different forms such as 30 (rounds), 6-man (patrol), and 4th (Division). There is a need to recognize these distinct forms and also to provide some means of tagging them within the proposed framework.

The specific problems which appear in the selected messages for the study were the map coordinates, 4 or 6 digits, times in 24 hour notation, and the numerical designations for military units. Documents from another environment would no doubt exhibit other peculiarities.

Several different approaches have been tried but none so far has resulted in a format consistent with the present system. The 24 hour clock introduces a modular concept and the coordinates are scale dependent. In addition, the areas of interest are not necessarily nice neat rectangles and would complicate the description of given geographical regiona.

We suggest the solution of introducing a fourth section to the dictionary specifications under the heading "RANGE." Each word which might have a numerical value associated with it would also have a corresponding entry as, for example, with respect to time:

Example: RANGE (HOURS,3)

0401-1200: MORNING.

1201-2000: AFTERNOON

2001-0400: EVENING

RANGE is a special code word, HOURS refers to the word in the dictionary and 3 specifies the number of entries in the table. MORNING, AFTERNOON and EVENING are concept names.

Corresponding to this declaration is an expanded expression to define the word "HOURS", using two additional control words.

Example: HOURS: IF COMP (-5-1)

THEN RANGE

ELSE

COMP is a special function which will return a true or false result after testing for the presence of a numerical word in the preceding 5 words. If the answer is true then execute RANGE - meaning find the table corresponding to HOURS and compare number with entries. If the number falls within the limits assign the corresponding concept name.

The RANGE directive must always be preceded by the COMP test, which finds and saves the pointer to the numerical string.

Further study is needed to accommodate other number representations in addition to the problems mentioned above. Combinations of hand tagging, character manipulation, and extracting capabilities should all be explored to come up with a generalized method rather than having to create additional functions for each new application.